IEEE S&P · 2026

Who *Taught* the Lie?

Responsibility Attribution for Poisoned Knowledge in Retrieval-Augmented Generation

Baolei Zhang†* Haoran Xin†* Yuxi Chen‡ Zhuqing Liu§ Biao Yi† Tong Li† Lihai Nie† Zheli Liu† Minghong Fang¶

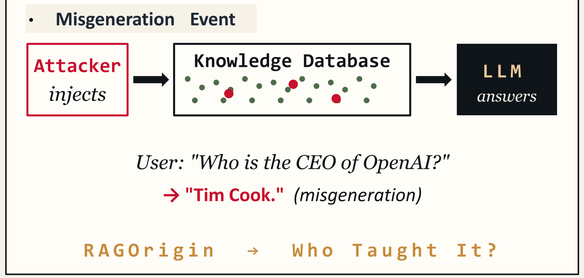
*†Nankai University ‡Independent Researcher §University of North Texas ¶University of Louisville *Equal contribution*

01 The Problem

> *A RAG system tells a user that **Tim Cook** is the CEO of OpenAI. **Why?** Someone in the knowledge database taught it that lie.*

RAG systems retrieve external texts to ground LLM answers. But an attacker can inject a handful of poisoned passages into the knowledge database, sometimes **<0.0002%** of the corpus, and steer the model toward an attacker-chosen response.

Defenses keep being bypassed by adaptive attacks. So we ask a different question: ***which texts caused this misgeneration?***



Key Observation

For a poisoning attack to succeed, the poisoned text must rank highly in retrieval AND push the LLM toward the target answer.

Both fingerprints are visible, even when the attacker is hiding.

Why Existing Work Falls Short

- ✗ **Forensics methods** assume access to gradients or parameters, infeasible for closed-source RAG.
- ✗ **RAG defenses** either fail to distinguish poisoned texts or assume more-than-half-clean retrieval.
- ✗ **RAGForensics** relies on a single LLM-judged semantic link, defeated when alignment is obfuscated.

At a Glance

<0.0002% → causes → ***0.99 ASR***

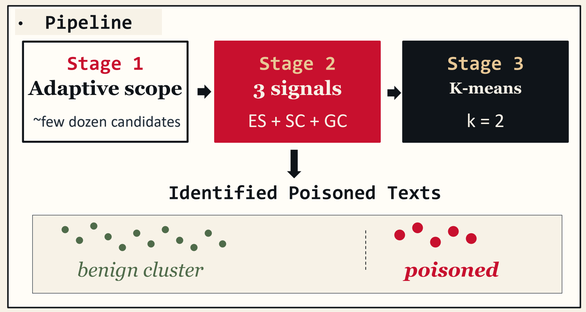
of corpus poisoned *attack success rate*

RAGOrigin cleans it up: DACC ≥ 0.97 with FPR, FNR ≤ 0.03

02 The Method · RAGOrigin

A black-box, three-signal forensic pipeline.

Given a misgeneration event ε = (q, r), RAGOrigin narrows the search to a few dozen candidate texts, scores each on three complementary axes, then clusters poisoned from benign, **no labels, no thresholds, no model internals.**



Embedding Similarity — ES

Does the text rank close to the question in the retriever's embedding space?

$$ES(u) = sim(E(q), E(u))$$

Semantic Correlation — SC

Given the text as context, how confidently does a proxy LLM predict the question's tokens?

$$SC(u) = avg\ log\ P(q \mid u)$$

Generation Influence — GC

How likely is the proxy LLM to produce the incorrect response with just this text as context?

$$GC(u) = avg\ log\ P(r \mid q, u)$$

Combined Responsibility Score

$$RS(u) = ES(u) + SC(u) + GC(u) / 3$$

z-normalized · averaged · clustered by K-means (k=2)

Algorithm · Black-box Workflow

01	**Receive event**	*ε = (q, r) from user report*
02	**Adaptive scope**	*expand top-k until coverage stable*
03	**Score candidates**	*compute ES + SC + GC for each text u*
04	**Cluster & flag**	*K-means(scores, k=2) → poisoned set*

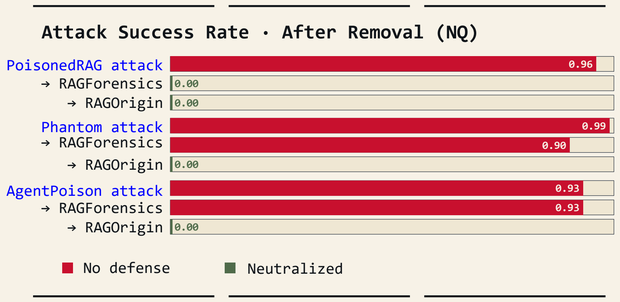
03 Results

Outperforms 8 baselines on 7 datasets and 15 attacks.

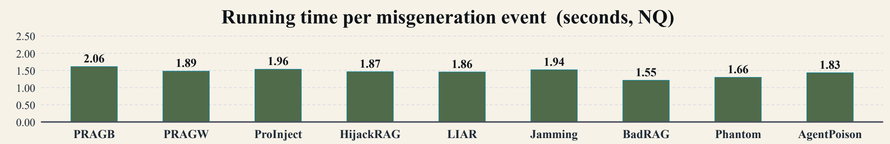
Across NQ, HotpotQA, MS-MARCO, BoolQ, SQuAD, ELI5, and a 16.7M-text corpus, RAGOrigin sustains detection accuracy **≥ 0.97** with false positive and false negative rates **≤ 0.03.**

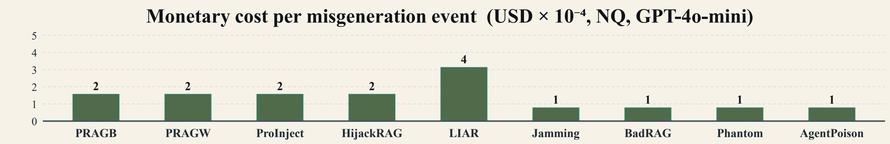
Defense	DACC ↑	FPR ↓	FNR ↓
Norm	0.62	0.16	0.77
PPL	0.55	0.00	1.00
SELF-RAG	0.64	0.49	0.12
RobustRAG	0.86	0.18	0.09
PFDNN	0.81	0.33	0.02
RAGForensics	0.69	0.03	0.44
RAGOrigin	**0.99**	**0.01**	**0.00**

Averaged across 9 attacks on NQ. Full per-attack tables in §6.2.



RAGOrigin is fast and cheap to run per misgeneration event





04 Holds Under Pressure

Robust to adaptive attacks and noisy real-world conditions.

Adaptive · 1	Adaptive · 2	Adaptive · 3
Benign-text perturbation	**Poisoned-text perturbation**	**Adversarial Hotflip**
DACC ≥ 0.98	DACC ≥ 0.98	DACC ≥ 0.97

Attackers given full knowledge of RAGOrigin's design (§6.4).

- ✓ **Paraphrased questions** — DACC stays ≥ 0.98.
- ✓ **Multi-attacker scenarios** — multiple adversaries on same/different questions.
- ✓ **Multi-hop & prompt-injection** — chained-text attacks. DACC ≥ 0.96.
- ✓ **Dynamic knowledge database** — corpus changes after misgeneration.
- ✓ **Noisy user reports** — paraphrased responses recovered correctly.
- ✓ **Open-ended QA (ELI5)** — DACC = 1.00 across all 9 attacks.
- ✓ **Scales to 16.7M texts** — drop-in with no degradation.
- ✓ **Robust across retrievers & LLMs** — Contriever, E5, GPT-4o, Llama-3

Contributions

① Novel black-box responsibility attribution framework for RAG poisoning.

② Multi-signal scoring + adaptive scope determination.

③ Robust evaluation across 15 attacks, 7 datasets, 3 adaptive strategies.

Limitations & Future Work

- Assumes honestly-reported misgeneration events; false-flag attacks need human-in-the-loop.
- Extension to multi-turn dialogue and multimodal RAG.
- Untargeted poisoning attribution remains open.

Contact: minghong.fang@louisville.edu